DOCUMENT RESUME

ED 479 334 TM 035 138

AUTHOR	Luppescu, Stuart
TITLE	DIF Detection in HLM.
PUB DATE	2002-04-08
NOTE	15p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Item Bias; Item Response Theory; Simulation; Test Items
IDENTIFIERS	*Hierarchical Linear Modeling; Item Bias Detection; Rasch Model; Root Mean Square (Statistics)

ABSTRACT

This study compared the ability of hierarchical linear modeling (HLM) to detect differential item functioning (DIF) to standard DIF detection methods, such as Rasch difficulty difference. The big advantages to using HLM for DIF detection are that the person abilities so produced are adjusted for any DIF in the items, and the DIF can then be modeled as a function of other predictors at a lower level in the same analysis. Data were simulated, and 180 data sets were produced to compare DIF detection approaches. In most cases, the amount of root mean square error (rmse) for Rasch and HLM is similar, but consistent points of difference occur when the number of people is small and the proportion of people in the focal group is low. In that case, the HLM rmse is larger, and when the number of people is large, the size of the DIF is small, and the proportion of people in the focal group is small, the HLM rmse is smaller. The paper discusses reasons HLM can produce better estimates than conventional methods in some circumstances. (SLD)

DIF Detection in HLM*

Stuart Luppescu

April 8, 2002

1 Introduction

Hierarchical linear models with discrete outcomes (Bernoulli, binomial, categorical (ordered and multinomial), and Poisson count) have been possible since the introduction of HGLM (hierarchical generalized linear models) several years ago. The extension of HGLM to IRT-style item analysis was a natural progression. The details of such analyses was outlined in a recent *JEM* article by Akihiko Kamata (Kamata, 2001) and in the HLM textbook (Raudenbush and Bryk, 2002). This little study extends Kamata's framework to include DIF detection. Its purpose is to compare the ability of HLM to detect DIF to standard DIF detection methods such as Rasch difficulty difference. Two big advantages to using HLM for DIF detection are

1. the person abilities so produced are adjusted for any DIF in the items; and
2. the DIF can then be modeled as a function of other predictors at a lower level in the same analysis.

2 Definition of Terms

DIF Differential Item Function. An item being systematically easier (or more difficult) for members of a particular group because of the content or format of the item and background or cultural knowledge or some other characteristic of the group. Classic example: "Coxswain : Shell :: President : ?" is very difficult for people not familiar with crew. The result of DIF is item or test bias in favor of one group.

Focus Group The group that is being considered as the subject of DIF analysis. This could be women, minorities, immigrants, etc.

Reference Group The remainder of the population not in the focus group.

Mantel-Haenszel A common DIF detection procedure which relies on the ratio the odds of members of the focus group answering an item correctly compared to the odds of the reference group answering the item correctly, conditional on total raw score.

Rasch Difficulty Difference A Rasch-based DIF detection method. Defined as the difference in item difficulty conditional on person ability. It has been shown to be computationally equivalent to Mantel-Haenszel (Schulz et al., 1996).

*Paper presented at the AERA Annual Meeting, April 2002, New Orleans

3 Procedure

3.1 Simulation

The data for this study were simulated using SIMTEST software (Luppescu, 2000). One hundred eighty data sets were produced: five each in 36 ($3 \times 4 \times 3$) different conditions.

Number of people	100	250	500	
Amount of DIF (logits)	0.25	0.50	0.75	1.0
Fraction of people in focus group	10%	25%	50%	

Each data set had 50 items. Of the 50 items, every 5th item (5th, 10th, 15th, etc.) was simulated to contain DIF. The individual responses were simulated using the 3-parameter logistic model:

$$p(x=1) = c - (1-c)\frac{exp(a(b-(d+DIF))}{1+exp(a(b-(d+DIF)))}$$

where
b=person ability $\sim N(0.5, 1)$
d=item difficulty $\sim N(0, 1)$
DIF=the DIF (0 for people and items without DIF)
a=discrimination or slope $\sim U(0.667, 1.5)$
c=lower asymptote or pseudo-guessing parameter $\sim U(0, 0.2)$

The generated probability $p(x=1)$ of each person's response to each item is compared to a uniform random number $\sim U(0,1)$. If the probability is greater than the random number, the response is assigned the value of 1; otherwise, 0.

3.2 Rasch DIF Detection

I have detailed the conventional method of DIF detection using Rasch previously (Luppescu, 1993). In short, all the items are calibrated on all the people. The person abilities from this first run are then used as anchors in subsequent runs calibrating the focal and reference groups separately. The differences in item difficulties from the latter two runs gives the estimated DIF.

4 HLM DIF Detection

In item analysis in HLM, individual item responses are entered as the outcomes, and a set of dummies indicate the item to which the response belongs. These dummies are effects coded with the intercept indicated by -1, and the item the response is for coded 1. Here is a section of the level-1 file:

person	ditem2	ditem3	ditem4	ditem5	ditem6	ditem7	ditem8	ditem9	ditem10	ditem11	ditem12	ditem13	ditem14	ditem15	ditem16	y
Person00001	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1

```
Person00001  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0 1
Person00001  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0 1
Person00001  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0 1
Person00001  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0 0
Person00001  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0 1
Person00001  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0 1
Person00001  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0 0
Person00001  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0 1
Person00001  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0 1
```

The level-2 file consists of the person ID, and a dummy indicating focal group membership.

```
 person       focus

Person00001     1
Person00002     1
Person00003     1
Person00004     1
Person00005     1
Person00006     1
Person00007     1
Person00008     1
Person00009     1
Person00010     1
Person00011     0
Person00012     0
Person00013     0
Person00014     0
Person00015     0
Person00016     0
Person00017     0
Person00018     0
Person00019     0
Person00020     0
```

The HLM model looks like this:

```
Level-1 Model

        Prob(Y=1|B) = P

        log[P/(1-P)] = B0 + B1*(DITEM2) + B2*(DITEM3) + B3*(DITEM4) +
B4*(DITEM5) + B5*(DITEM6) + B6*(DITEM7) + B7*(DITEM8) +
B8*(DITEM9) + B9*(DITEM10) + B10*(DITEM11) + B11*(DITEM12) +
B12*(DITEM13) + B13*(DITEM14) + B14*(DITEM15) + ... + B49*(DITEM50)

Level-2 Model
        B0 = G00 + U0
        B1 = G10 + G11*(FOCUS)
        B2 = G20 + G21*(FOCUS)
        B3 = G30 + G31*(FOCUS)
        B4 = G40 + G41*(FOCUS)
        B5 = G50 + G51*(FOCUS)
        B6 = G60 + G61*(FOCUS)
        B7 = G70 + G71*(FOCUS)
        B8 = G80 + G81*(FOCUS)
        B9 = G90 + G91*(FOCUS)
        B10 = G100 + G101*(FOCUS)
        B11 = G110 + G111*(FOCUS)
```

```
B12 = G120 + G121*(FOCUS)
B13 = G130 + G131*(FOCUS)
B14 = G140 + G141*(FOCUS)
...
B49 = G490 + G491*(FOCUS)
```

All the item focal group dummies are grand-mean centered and fixed, so the intercept will be 0 and is the mean of the item difficulties. (This is necessary to resolve the indeterminacy of scale problem common to all IRT models.) The fixed effects for the item dummies become the item difficulties, and the random effect on the intercept is the person ability. The coefficient for each of the focal group dummies is the amount of DIF for that item.

Here is the fixed effects output table for one such run:

Final estimation of fixed effects: (Population-average model)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	0.510260	0.091843	5.556	99	0.000
For DITEM2 slope, B1					
INTRCPT2, G10	1.829907	1.293485	1.415	4901	0.157
FOCUS, G11	6.472844	12.670030	0.511	4901	0.609
For DITEM3 slope, B2					
INTRCPT2, G20	1.175504	0.275524	4.266	4901	0.000
FOCUS, G21	-0.943093	0.788108	-1.197	4901	0.232
For DITEM4 slope, B3					
INTRCPT2, G30	1.133919	0.282696	4.011	4901	0.000
FOCUS, G31	-2.288452	0.750152	-3.051	4901	0.003
For DITEM5 slope, B4					
INTRCPT2, G40	1.069872	1.283261	0.834	4901	0.405
FOCUS, G41	7.317327	12.669668	0.578	4901	0.563
For DITEM6 slope, B5					
INTRCPT2, G50	1.006696	0.258698	3.891	4901	0.000
FOCUS, G51	-0.156994	0.869540	-0.181	4901	0.857
For DITEM7 slope, B6					
INTRCPT2, G60	0.645075	0.234003	2.757	4901	0.006
FOCUS, G61	-0.353727	0.770810	-0.459	4901	0.646
For DITEM8 slope, B7					
INTRCPT2, G70	0.760882	0.241497	3.151	4901	0.002
FOCUS, G71	-0.482401	0.773751	-0.623	4901	0.533
For DITEM9 slope, B8					
INTRCPT2, G80	1.313028	0.286437	4.584	4901	0.000
FOCUS, G81	-0.497363	0.880548	-0.565	4901	0.572
For DITEM10 slope, B9					
INTRCPT2, G90	0.244496	0.223715	1.093	4901	0.275
FOCUS, G91	1.589476	1.086608	1.463	4901	0.144
For DITEM11 slope, B10					
INTRCPT2, G100	0.385138	0.220180	1.749	4901	0.080
FOCUS, G101	-0.064908	0.765602	-0.085	4901	0.933
For DITEM12 slope, B11					
INTRCPT2, G110	0.519969	0.233412	2.228	4901	0.026
FOCUS, G111	1.283395	1.089200	1.178	4901	0.239
For DITEM13 slope, B12					
INTRCPT2, G120	0.023522	0.207539	0.113	4901	0.910
FOCUS, G121	0.336888	0.761088	0.443	4901	0.658
For DITEM14 slope, B13					
INTRCPT2, G130	0.434392	0.222488	1.952	4901	0.051

```
        FOCUS, G131         -0.119635   0.766452   -0.156      4901     0.876
For  DITEM15 slope, B14
     INTRCPT2, G140          0.492561   0.226372    2.176      4901     0.029
        FOCUS, G141         -0.674887   0.726740   -0.929      4901     0.353
For  DITEM16 slope, B15
     INTRCPT2, G150          0.532513   0.237548    2.242      4901     0.025
        FOCUS, G151         -2.111216   0.771293   -2.737      4901     0.007
For  DITEM17 slope, B16
     INTRCPT2, G160          0.895712   0.253623    3.532      4901     0.001
        FOCUS, G161          0.865902   1.094933    0.791      4901     0.429
For  DITEM18 slope, B17
     INTRCPT2, G170          0.698943   0.237152    2.947      4901     0.004
        FOCUS, G171          0.184954   0.861678    0.215      4901     0.830
For  DITEM19 slope, B18
     INTRCPT2, G180          0.452027   0.225958    2.000      4901     0.045
        FOCUS, G181         -1.080230   0.713754   -1.513      4901     0.130
For  DITEM20 slope, B19
     INTRCPT2, G190         -0.146289   0.203350   -0.719      4901     0.472
        FOCUS, G191          0.034946   0.718010    0.049      4901     0.962
For  DITEM21 slope, B20
     INTRCPT2, G200          0.340982   0.218761    1.559      4901     0.119
        FOCUS, G201         -0.506467   0.723762   -0.700      4901     0.484
For  DITEM22 slope, B21
     INTRCPT2, G210          0.109468   0.209748    0.522      4901     0.601
        FOCUS, G211         -0.249228   0.720353   -0.346      4901     0.729
For  DITEM23 slope, B22
     INTRCPT2, G220          0.109665   0.209875    0.523      4901     0.601
        FOCUS, G221          0.241173   0.761904    0.317      4901     0.751
For  DITEM24 slope, B23
     INTRCPT2, G230         -0.268518   0.202613   -1.325      4901     0.185
        FOCUS, G231         -0.730189   0.716366   -1.019      4901     0.309
For  DITEM25 slope, B24
     INTRCPT2, G240          0.153624   0.211228    0.727      4901     0.467
        FOCUS, G241          0.192330   0.762380    0.252      4901     0.801
```

The value of the intercept, 0.51, agrees with the average person ability of the generating parameters, 0.50. The coefficients for INTRCPT2 are the item difficulties. In HLM, the log odds of a correct response is the *sum* of the item difficulty and the person random effect, while in conventional IRT approaches, the log odds of a correct response is the *difference* of the item difficulty and person ability. For this reason, the coefficients in the fixed effects table correspond to (-1) times the item difficulty one would get from conventional IRT analysis.

The items with DIF in this section of the output are 5, 10, 15, 20, and 25. The coefficients for G41, G91, G141, G191, and G241 are the estimated DIF for those items. Note that G41 and G91 are much larger than the others, indicating a strong probability of DIF.

5 Results

In this section I compare the results from the two methods of DIF detection: Rasch difficulty difference, and HLM. For each of the 36 combinations of generating parameters, there were five data sets simulated. For each item, the HLM DIF, taken from the fixed effects tables, and the Rasch DIF (the differences in item difficulties for focal and reference groups) were calculated, and the root mean squared error for the focal group and the reference group within each of the 36 combinations calculated according to this formula:

$$rmse = \sqrt{\frac{\sum (DIF_{est} - DIF_{act})^2}{nj}}$$

where
DIF_{est} is the estimated DIF (either Rasch or HLM)
DIF_{act} is the actual DIF (from the generating parameters)
n is the number of items (50)
j is the number of data sets within each combination (5)

The following plots show the rmse for each of the 36 combinations of generating parameters.

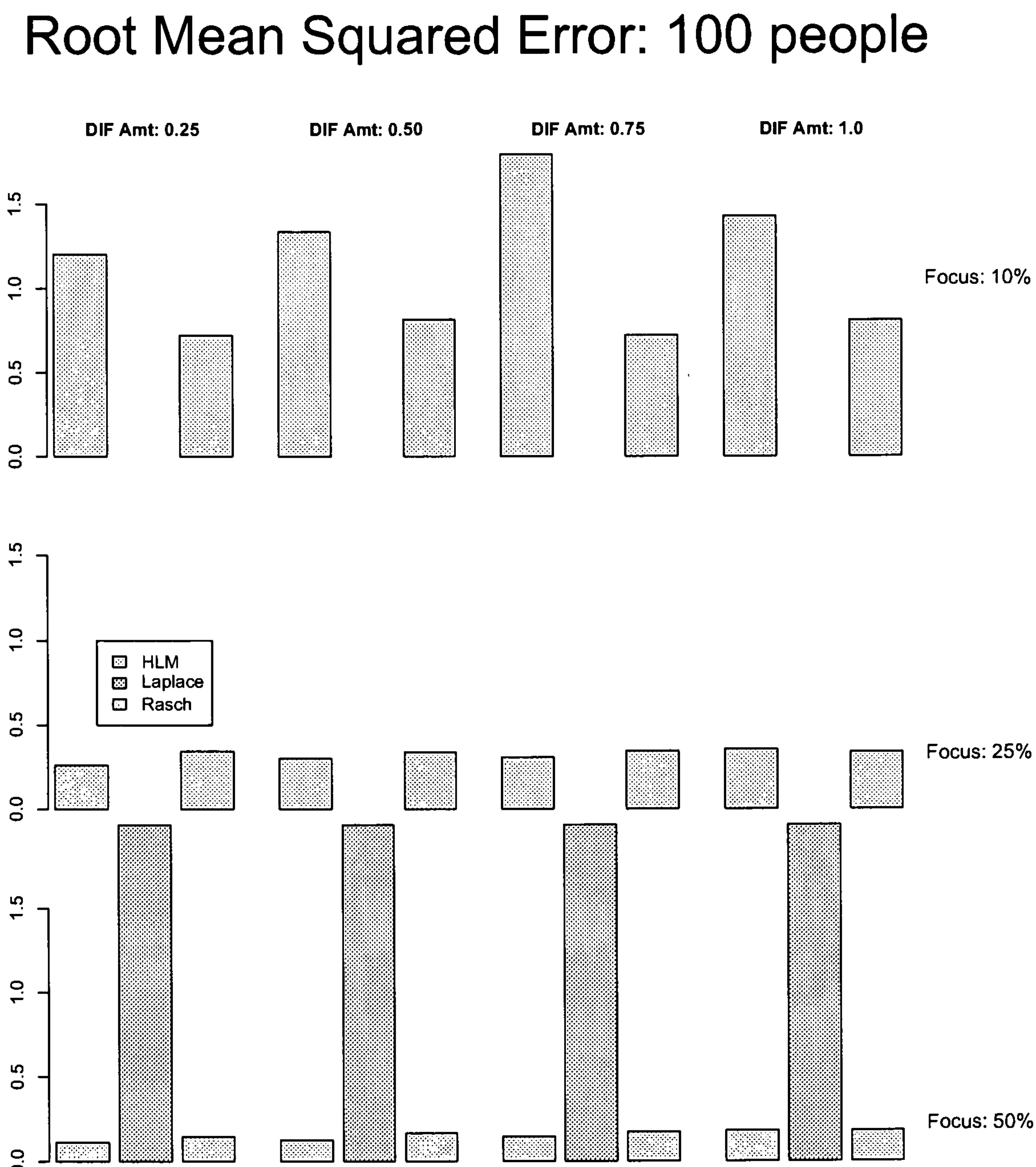

Figure 1: Root mean squared error for 100-person data sets

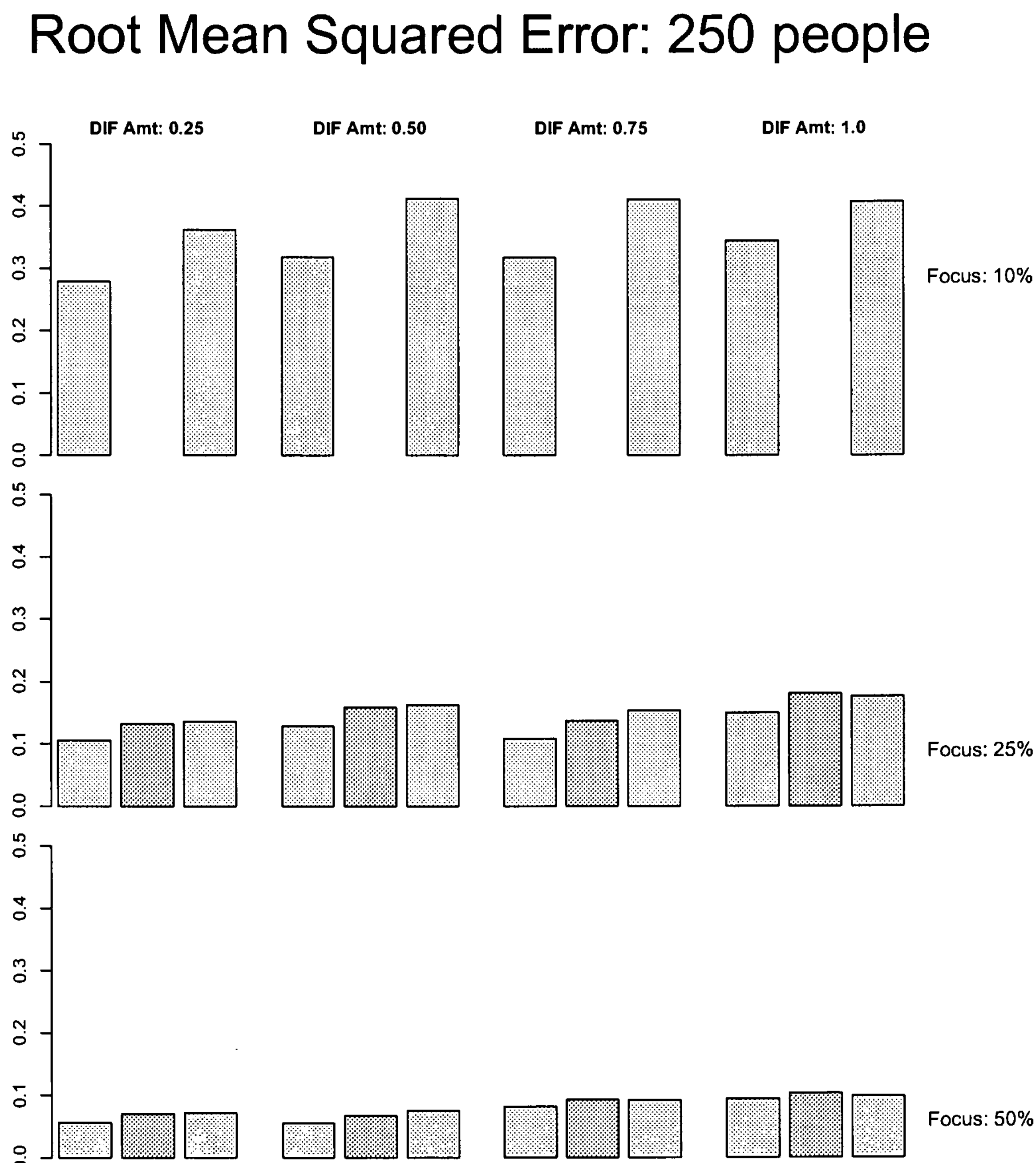

Figure 2: Root mean squared error for 250-person data sets

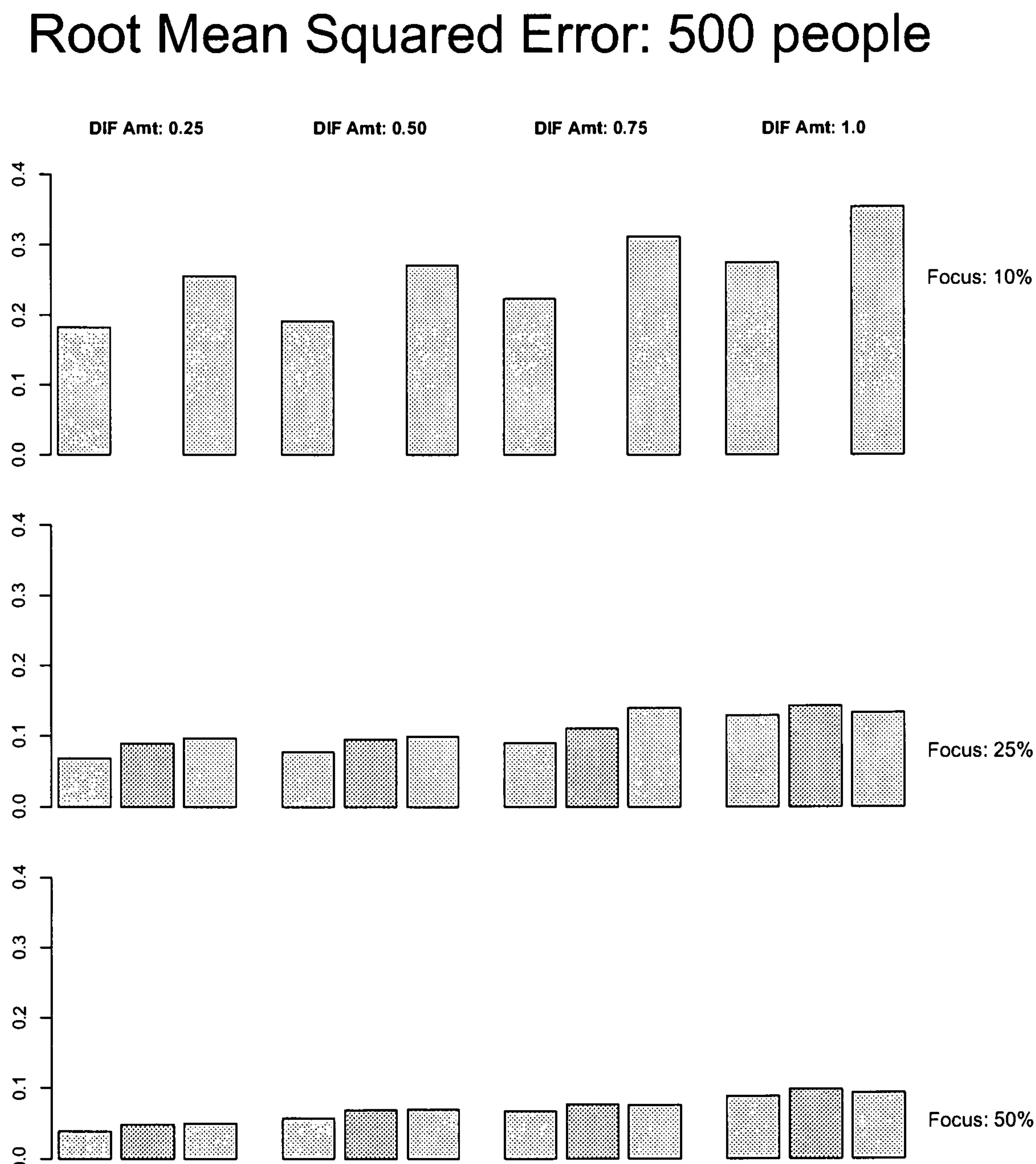

Figure 3: Root mean squared error for 500-person data sets

In most cases, the amount of rmse for Rasch and HLM is similar. Consistent points of difference are when the number of people is small and the proportion of people in the focal group is low, in which case the HLM rmse is larger; and when the number of people is large, the size of the DIF is small, and the proportion of people in the focal group is small, in which case the HLM rmse is smaller. The following box plots illustrate two cases when the HLM and Rasch rmse are different.

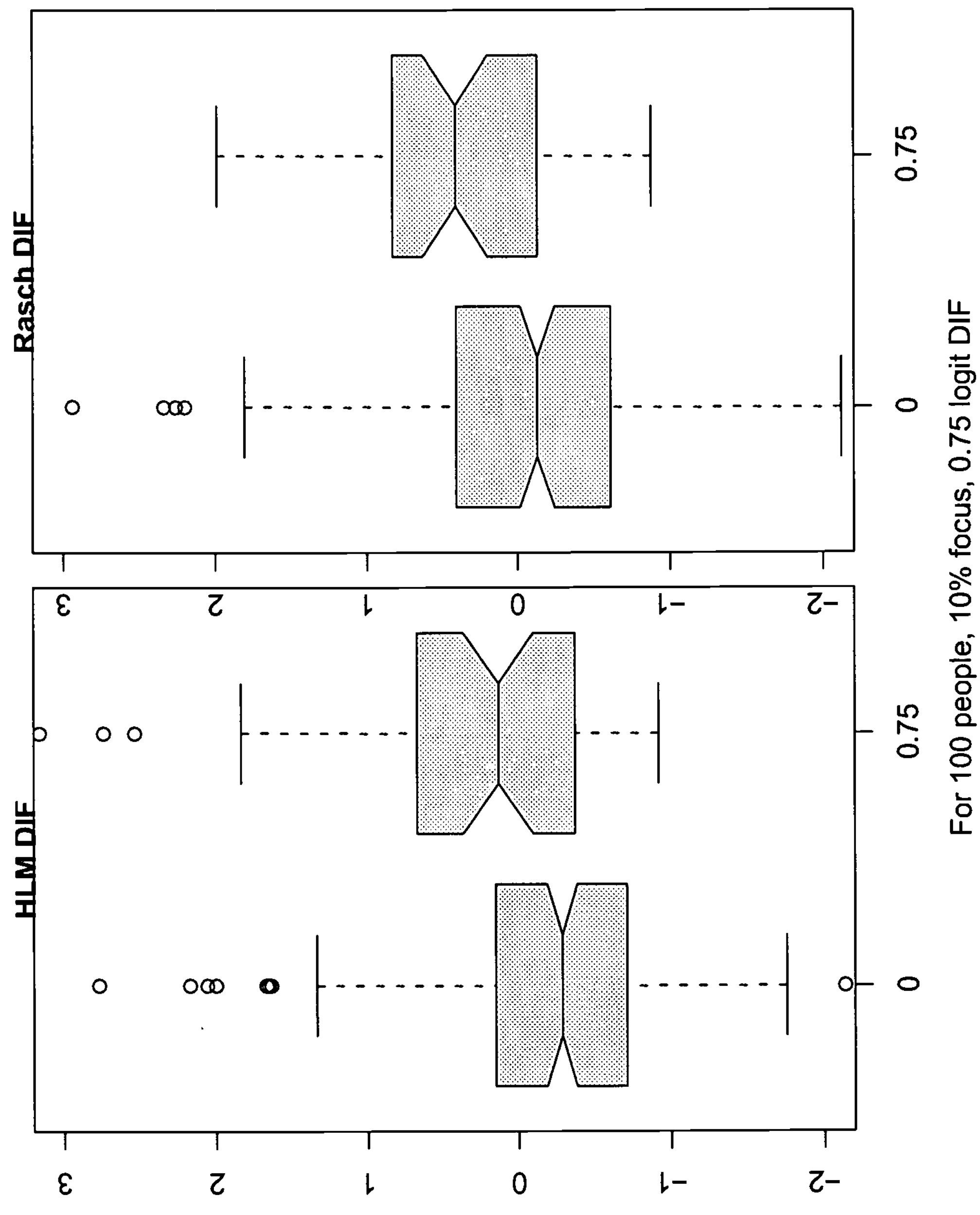

Figure 4: For 100 people, 10% focus, 0.75 logit DIF

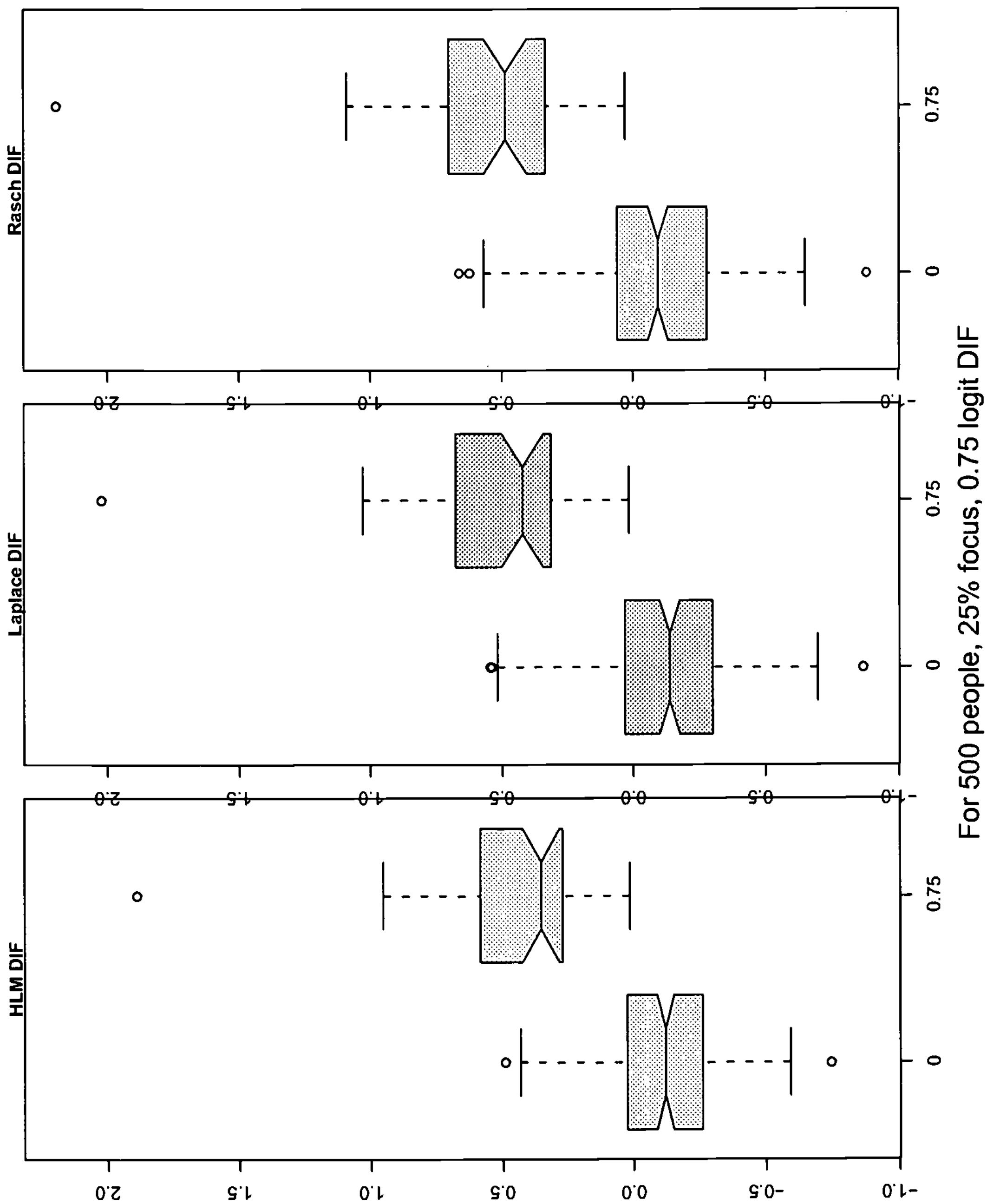

Figure 5: For 500 people, 25% focus, 0.75 logit DIF

What you can't see in left side of figure 4 is that there about 3 or 4 points off the scale of the graph at about 5, 6, and 14. All of the outliers are cases in which all of the members of the focal group got the item correct. Because the BIGSTEPS (Linacre and Wright, 2000) estimation procedure is unable to produce parameter estimates for extreme case, it uses some sort of Bayesian technique to assign measures to extreme cases. The algorithm used by HLM can produce parameter estimates, but these are likely to have rather extreme values. When the focal group is small, extreme cases are much more likely to occur; this explains why we see larger rmse for HLM in the cases with a small number of people, and a small fraction of people in the focal group.

In virtually all of the remaining cases, HLM does as well as or better than the conventional DIF detection methods. The box plots in figure 5 show the distribution of DIF for the 500-person data set, where 25% of the people are in the focal group, with 0.75 logits DIF. While the box plots all look rather similar, the distribution of DIF is a bit tighter for the HLM-estimated DIF, and slightly less so for Laplace and Rasch.

The darker blue bars represent the average root mean squared error of the DIF estimation using the Laplace approximation to maximum likelihood in HLM. In 91 out of 180 cases, HLM was not able to produce Laplace estimates. The program terminated with messages saying the H matrix was not invertible, or that the deviance after a certain iteration was Not A Number. This is due to the Laplace transform using Fischer scoring, which sometimes produces estimates outside the parameter space, causing the program to blow up. This is most likely to happen when variances are close to zero, which is more likely to happen when sample sizes are small.

The question of why HLM produces better estimates than conventional methods has no definite answer. I do have two theories:

1. Rasch methods require two calibrations to estimate the two sets of item difficulties that are being compared. It could be that there is additional error involved in calculating two estimates and taking the difference, than in estimating the DIF directly in a single run (as HLM does).

2. HLM produces estimates for the gammas based on its variance-covariance matrix. The EM algorithm may be a more efficient estimator of the the variance-covariance matrix than the one BIGSTEPS employs.

In addition, I have received a valuable comment from Steve Raudenbush regarding this. Steve's idea is this: Bigsteps estimates a fixed effect for each person and each item, while HLM estimates a fixed effect for each item and a single random effect for all the people. HLM's parameter estimates may be better because HLM has to estimate fewer parameters.

References

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38(1):79–93.

Linacre, J. M. and Wright, B. D. (2000). Bigsteps. Computer Program.

Luppescu, S. (1993). Dif detection examined. *Rasch Measurement Transactions*, 7(2):285.

Luppescu, S. (2000). Simtest. Computer Program.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Thousand Oaks, CA, second edition.

Schulz, E. M., Perlman, C. L., Rice, W. K., and Wright, B. D. (1996). An empirical comparison of rasch and mantel-haenszel procedures for assessing differential item functioning. In Englehard, G. and Monsaas, J., editors, *Objective Measurement: Theory into Practice, 3*, pages 65–82. Albex Publishing Corporation, Norwood, NJ.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

TM035138

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: DIF Destection in HLM

Author(s): Stuart Luppescu

Corporate Source: Univ. of Chicago

Publication Date: Apr 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: Stuart Luppescu	
Organization/Address: Univ. of Chicago 1313 E 60th Street Chicago, IL 60637	Telephone: 773-702-3186	FAX: 773 702-2010
	E-Mail Address: s-luppescu@uchicago.edu	Date: 4/4/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com